Florence, Italy 12-16 July 2015
ICSV22
The 22nd International Congress on Sound and Vibration

THE ACUITY OF COLOURATION PERCEPTION USING NON-INDIVIDUALISED DYNAMIC BINAURAL SYNTHESIS

Darius Satongar and Yiu Wai Lam
University of Salford, Salford, UK
e-mail: dsatongar1@edu.salford.ac.uk

Chris Pike
BBC Research and Development, Salford Quays, Salford, UK

Spectral inaccuracies in binaural synthesis caused by factors such as non-individualised HRTFs and headphone compensation filters can induce colouration artifacts; changing the perception of an intended auditory event. Detection thresholds for induced colouration artifacts can be used to approximate colouration acuity. If the colouration acuity with binaurally synthesised sound sources is perceptually equivalent to that with real sound sources, the binaural synthesis system can be considered acceptable for evaluating colouration in synthesised environments. Perceptual colouration detection thresholds using both real and binaurally synthesised sound sources were measured and compared as an indicator of colouration acuity.

1. Introduction

The perceptual assessment of spatial audio reproduction systems commonly focuses on the central listening position (CLP) or *sweet spot*. This position can be identified as the best listening position [1] and is usually geometrically equidistant from all reproduction loudspeakers [2]. However, in domestic listening environments many listeners will not be seated centrally within the loudspeaker array and the sweet spot can be small [3]. As an approach to consider spatial audio reproduction systems in realistic listening environments it is important to consider perceptual assessment at multiple listening positions. Achieving blind, subjective comparisons of the effect of different listening positions is problematic when using in-situ methods due to physical and logistical factors. Some methods have been implemented by [4][5][6][7] commonly taking two or sometimes a small selection of listening positions as an independent variable. Reproducing the perception at different listening positions binaurally (either by simulated or measured cues) has also been implemented in [2] and for the application of Wave Field Synthesis reproduction, a dynamic binaural system has been implemented in [8][9], where it was found that localisation acuity with binaural simulation of single loudspeakers was comparable to real loudspeakers in a localisation task.

The ability to detect colouration artifacts using a binaural simulation has so far only been considered to a limited extent [10]. This fact was also highlighted by [11] who implemented a test for changes in colouration in WFS systems. Olive and Shuck [12] conducted tests assessing the sound

quality preferences of different loudspeakers for both in-situ and non-dynamic binaural simulation. Comparing results between in-situ and binaural simulation highlighted 'remarkably good agreement'. However, results showed that they achieved a higher number of significant ($p \leq 0.1$) factors and interactions when using the binaural system. The reason for this was not investigated directly. The effects of non-dynamic binaural simulation including HRTF personalisation was also investigated by Hiekkanen, Mäkivirta and Karjalainen [13]. Their study looked at a selection of attributes covering spatial and timbral domains. Results highlighted stimulus dependence but that individually equalised artificial head responses were acceptable for binaural synthesis of stereo loudspeakers. The most recent consideration of the problem [14] presented results for the differences in magnitude spectra between two different dummy heads using a number of spatial audio reproduction systems to simulate the experience of a non-individualised AVE. Results showed that although the differences in magnitude spectra were significant (up to 15dB for high frequencies) the effect was relatively consistent across all WFS systems and directions tested. The study concluded that the magnitude difference was a linear effect and the AVE system was later implemented for the subjective evaluation of colouration.

In this study we present results for the validation of a non-individualised dynamic Auditory Virtual Environment (AVE) to simulate CLP and non-CLP colouration artifacts using a colouration detection threshold (CDT) test.

2. Auditory Virtual Environment

An auditory virtual environment (AVE) has been developed using the spatially-sampled binaural room impulse response dataset [15] to simulate loudspeaker-based spatial audio reproduction in an existing auditory environment. The purpose of this simulation is to allow testing of various domestic spatial audio reproduction methods at multiple listening positions in a direct blind comparison. The AVE consists of a head-azimuth tracking system utilising infrared cameras to identify the position of reflective passive markers; this allows for accuracy of <0.1º in orientation. Tracking data is sent to a modified version of the open-source SoundScape Renderer [16] via TCP/IP which handles the real-time filter convolution. BRIR processing was modified to only dynamically change an initial region of the left and right impulse responses under head-rotations[17]. After removing leading silence that was consistent across BRIRs, 50ms (relative to BRIR start, not onset) was chosen as the mixing-time to allow for dynamic early reflections. The static region was extracted from the 0º head-azimuth BRIR. The loudspeaker input signal was sent using JACK from Max (https://cycling74.com) which was also used to control the test and record the data. The system set-up under similar test conditions [18] had a mean total system latency of 41.2ms ($\sigma = 2.6ms$). The non-individualised binaural room impulse responses were measured using a Brüel & Kjær (B&K) head-and-torso simulator (HATS). Headphone compensation filters were calculated using the method of [19] measured on the HATS using STAX SR-202 electrostatic headphones in which multiple measurements were made to account for differences in ear-to-headphone coupling. The comb-filtered signal was replayed from a real or binaurally simulated Genelec 8030A loudspeaker at 315º (front, right) of the listener at a 2.1m distance. An acoustically transparent curtain was placed at a 2m radius around the listener.

3. Colouration

Unlike localisation, colouration is not an objective, self-referencing metric but one that requires an explicit reference [20]. To use an auditory virtual environment (AVE) to measure colouration perception induced by non-central listening in loudspeaker-based spatial audio we must ensure that the limitations of the AVE do not have systematic effects that will invalidate the results. In this test we measure participants' CDTs using an in-situ loudspeaker and the same loudspeaker simulated using the described non-individualized dynamic AVE. Measurements of CDTs for harmonic cosine

noise were originally presented by [21] and further measurements can be found in [22], [23] and [20]. Reflection threshold experiments less specific to the percept of colouration have also been conducted for a variety of conditions, see [24] and [25] for examples.

A non-exhaustive list of AVE limitations influencing the ability to detect colouration is:

1. Non-individualised BRIRs
2. Non-individualised headphone compensation filtering
3. Non-individualized binaural cues specifically influencing the binaural decolouration process (ITD, ILD, IACC)
4. Discretisation of dynamic cues and lack of translation, tilt and roll

This validation stage aims to verify that colouration artifacts observed in loudspeaker reproduction at different listening positions will be perceivable when simulated with the AVE and the difference in CDTs measured using the in-situ and AVE auralisation methods are small enough to be considered perceptually equivalent. Artifacts can be approximated by comb filtering at non-central listening positions with direct transmission path differences between loudspeakers in the region of 0-7ms; these delays can be approximated using simple geometric calculations.

4. Artificially Induced Colouration

Before each step both reference and coloured signals were played to the participant. The reference for the test was the original (uncoloured) white noise signal with a uniformly distributed power density function. For the coloured signal, comb filtering was artificially introduced using a harmonic cosine noise signal. Salamons [20] defines this as cosine due to sinusoidal notches in the frequency response and harmonic as the repetition has 0º phase-shift. The frequency response notches can be calculated using eq. (1).

$$f_i = \frac{2i-1}{2T} \tag{1}$$

where f_i is the frequency for notch integer i, T is the delay.
Figure. 2 shows the block processing for this signal.

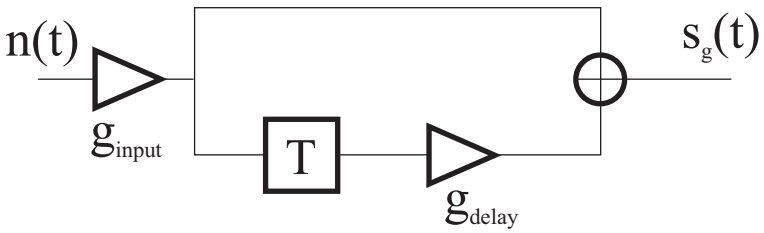

Figure 1. Block diagram for Harmonic Cosine Noise

Where T is the delay of the white noise signal, g_{delay} is the relative level of the delayed noise signal and therefore the level for which a threshold of detection is measured. g_{input} scales the noise signal to ensure that the output $s_g(t)$ remains a consistent level. For this experiment, $T=2ms$. The input signal n(t) was scaled by g_{input} using equation (2) to maintain level consistency between varying g_{delay}.

$$g_{input} = \frac{1}{\sqrt{1+g_{delay}{}^2}} \tag{2}$$

Looking at the effect of this filter in the frequency domain highlights linearly spaced notches with equal magnitude across the full frequency range. However, due to the spacing and bandwidth

of auditory filters, spectral notches are not as audible at higher frequencies as at lower frequencies. Passing the impulse response of the harmonic cosine noise generator through an auditory filter bank with ERB spaced gamma-tone filters using the Auditory Modelling Toolbox [26], Fig. 3 shows clear notches decreasing in severity in the upper frequency regions due to the auditory filter spacing. After a certain frequency limit the notches are imperceptible. The plot also shows that as the delay value T is increased, the fundamental notch (and therefore integer multiples therefore) lowers in frequency.

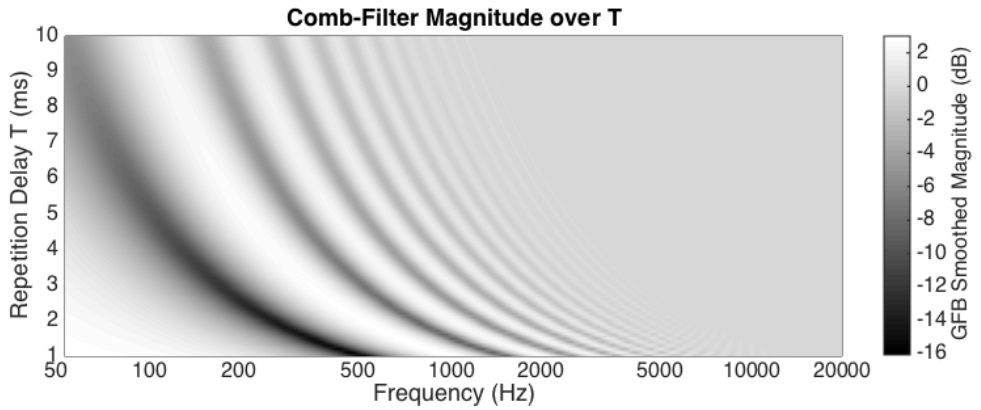

Figure 2. Gamma-tone filter bank smoothed power spectrum for harmonic cosine noise with different delay (T) values

Following direction in [20], a test using the top-down, 'adjustment method' was implemented to allow participants to find their own thresholds. The test was carried out in the University of Salford ITU-R BS.1116-1 conformant listening room. Participants reported their decision about perceived colouration using a Max graphical interface. The Max program also generated the coloured noise signals and recorded the experiment data.

Following a press of the 'play' button, two noise signals were played one after another; one uncoloured reference white noise signal and one coloured white noise signal using the colouration method described above. Noise signals had duration of 1s with a 750ms silence in-between. The order of reference/coloured was randomised for each play and the participants were informed of this. If a difference in colouration was perceived (answer YES) the amount of colouration was decreased in the next session, if no colouration is perceived (answer NO) the amount of colouration was increased. After the first reversal, the subject continued to move above and below their threshold until a green light on the GUI became red, this indicated 15 reversals had been made and the end of the test section. The initial step size of decreased colouration was randomly chosen from between 4dB to 8dB and was reduced with each step to a minimum of 1dB. Step sizes were not revealed to the participants.

Participants were given a training session before the test allowing them to become familiar with the interface, hear the effect of their decisions and understand the format of the test method. Participants with experience in listening tests were a prerequisite due to the reliance on finding their own threshold; this was assessed by a pre-test questionnaire. Each participant undertook two threshold tests for both auralisation methods giving a total of four CDT values per participant. The order of auralisation method presented to the participant was randomised between participants in either AABB or BBAA sequence.

5. Results

A total of 6 male listeners from the University of Salford undertook the experiment – some of which had used the AVE in a previous localisation test. Figure 4 shows the judgement responses for participant number three.

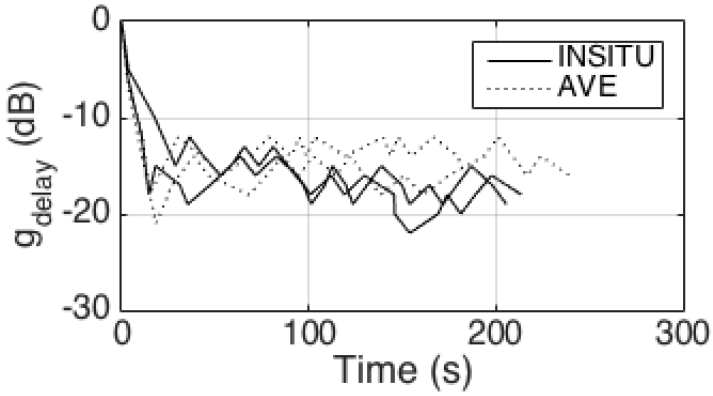

Figure 3. Real-time threshold judgement results for participant #3.

Although the adjustment method is efficient and usually less tiring for participants compared to paired comparison tests such as 2AFC, thresholds can sometimes be reported lower than their actual thresholds [20]. This could be caused by erroneously achieving a 'phantom threshold' in which a participants steps much farther beyond their actual threshold before the first reversal and never realign themselves; this is exacerbated by the reduction in step-size compared to the initial step-size meaning a realignment requires more 'NO' answers than the 'YES' answer they made to get there. The phantom thresholds were selected as measurements where the first reversal was below -30dB. -30dB was chosen due to it being unrealistically perceivable based on literature results. Phantom threshold values were measured in 3 of the 24 measurements (CDTs of -27.8 dB, -42.8 dB and -35.6 dB), twice for in situ and once for AVE. These judgements were consequently removed from further analysis. Figure 5 (left) shows the raw threshold values measured for each judgement (6 subjects, 2 systems, 2 repeats). Each participant is represented by a different symbol (+O*×□◇).

5.1 Analysis

From this data we can independently approximate the error induced by the AVE by creating an error sample – this secondary metric we call 'Δ Colouration Detection Threshold' (ΔCDT). For each subject, find all possible differences between the threshold values for the AVE and In-situ. The number of error judgements is equal to the number of repeats squared (N_{Δ} = 4 per subject). Performed across all subjects gives a sample of threshold errors for AVE auralisation. Note that non-trivial inter-subject variation is accounted for by only ever calculating the error per subject (i.e. never finding the threshold difference between AVE and in-situ across different subjects). This gave a ΔCDT sample size of N = 18, this value is reduced from the theoretical size of 24 due to removal of 3 phantom thresholds. Due to the small sample size a bootstrapped non-parametric estimate of the mean confidence intervals was created using 1000 repeated re-samples with replacement drawn from the raw ΔCDT sample. The mean value was used to estimate the central tendency due to the smaller number of cases and the increased problem of ties when using the median. Figure. 5 (right) shows the results.

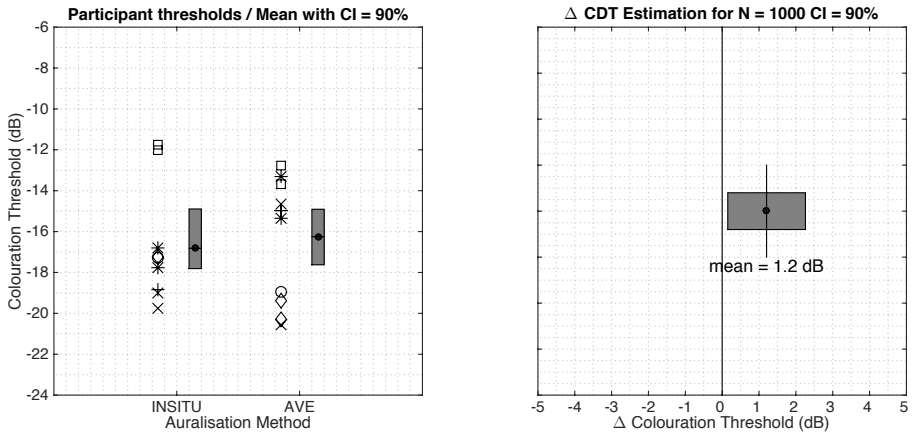

Figure 4. (left) Raw participant recorded colouration detection thresholds with bootstrapped median and 90% CIs. (right) ΔColouration Detection Threshold with 90% CIs

Salomons' [20] reported results for inter-subject variation is substantial; in a harmonic cosine threshold test with diotic presentation and $T = 2ms$, the range in participant reported CDTs was ≈6dB. We can see from measured data in Figure 5 (left) that inline with Salomons' studies there is considerable inter-subject variation in thresholds for both the in-situ and AVE. Grey boxes show bootstrapped 90% CIs for the variation in CDT mean. Comparing mean and non-parametric confidence intervals show good similarity. An interesting feature of the recorded thresholds is that for the AVE, data could indicate bimodal behaviour, however, without a larger number of participants we must assume a unimodal underlying distribution. If we now consider the effect of the AVE directly by using the ΔCT metric, Figure 5 (right) shows that there was a systematic increase in CDTs measured using the AVE and the lower 90% confidence interval is above 0dB, which indicates statistical significance. Looking at the influence of the AVE, we can see that the there was a mean systematic increase in CDT of 1.2dB.

6. Discussion

Mean colouration detection thresholds measured using both an in-situ and an AVE simulated loudspeaker were found to be very close when averaging over subjects. However, the large inter-subject variations in CDTs caused mean values for each auralisation method to be a poor parametric estimate of the influence of the AVE. When considering the systematic influence of the AVE using ΔCDT calculations we found that the AVE systematically increased the recorded CDT value by an average of 1.2dB. This signifies that when using an AVE, colouration acuity may be reduced. Therefore, careful consideration must be made when deciding to use a similar AVE to measure the perception of colouration artifacts. Although the systematic increase is not very large, perceptually small colouration artifacts could be unperceivable when using an AVE. However, when results for colouration are also averaged over subjects, the influence in colouration perception could also be considered insignificant due to the large variations in perception between subjects. The cause for the increase in CDTs could be due to the use non-perfect simulation of binaural cues by the AVE.

As a positive note on the plausibility of the AVE, although participants were explicitly told when reproduction was from the AVE or In-situ loudspeakers, 2 participants of the colouration test were convinced that they were listening to real speakers during the AVE sessions and had to remove the headphones to be convinced that they were not being fooled. Most participants felt favourably towards the AVE's plausibility.

7. Conclusions

A colouration detection threshold test found that the subjective variability in CDTs were non-trivial and mean values across participants were similar between In-situ and AVE auralisation methods. Using a secondary metric, a 1.2dB mean increase in measured CDT was found for the AVE. From this we can deduce that colouration acuity decreases by a small amount using an AVE. Although the systematic effect was statistically significant, the perceptual difference can be considered insignificant for amounts of colouration that are not close to the perceptual threshold. Inter-subject variation in reported thresholds was, however, non-trivial. This could possibly be due to the method of determining the threshold. More investigation is thus required, possible implementing different directions and repetition delays to quantify the effect more comprehensively.

REFERENCES

[1] F. Toole, *Sound Reproduction: Loudspeakers and Rooms*. Focal Press, 2008.

[2] N. Peters, "Developing Sound Spatialization Tools for Musical Applications with Emphasis on Sweet Spot and Off-Center Perception," McGill University, Montreal, Canada, 2010.

[3] B. S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial Sound With Loudspeakers and Its Perception: A Review of the Current State," *Proc. IEEE*, vol. 101, no. 9, pp. 1–19, 2013.

[4] E. Bates, G. Kearney, and D. Furlong, "Localization accuracy of advanced spatialisation techniques in small concert halls," *J. Acoust. Soc. Am.*, vol. 121, no. 5, pp. 3069–3070, 2007.

[5] P. Stitt, S. Bertet, and M. van Walstijn, "Off-Centre Localisation Performance of Ambisonics and HOA For Large and Small Loudspeaker Array Radii," *Acta Acust. united with Acust.*, vol. 100, no. 5, pp. 937–944, Sep. 2014.

[6] K. Hamasaki, T. Nishiguchi, R. Okumua, and Y. Nakayama, "Wide Listening Area with Exceptional Spatial Sound Quality of a 22.2 Multichannel Sound System," in *122nd Audio Engineering Society Convention*, 2007, pp. 1–22.

[7] R. Conetta, "Towards the automatic assessment of spatial quality in the reproduced sound environment," University of Surry, 2011.

[8] H. Wierstorf, A. Raake, and S. Spors, "Localization of a virtual point source within the listening area for Wave Field Synthesis," in *133rd Audio Engineering Society Convention*, 2012, vol. c, pp. 1–9.

[9] H. Wierstorf, S. Spors, and A. Raake, "Perception and evaluation of sound fields," in *59th Open Seminar on Acoustics*, 2012.

[10] H. Wittek, F. Rumsey, and G. Theile, "Perceptual Enhancement of Wavefield Synthesis by Stereophonic Means," *J. Audio Eng. Soc.*, vol. 55, no. 9, pp. 723–751, 2007.

[11] H. Wierstorf, C. Hohnerlein, S. Spors, and A. Raake, "Coloration in Wave Field Synthesis," in *55th AES International Conference*, 2014, pp. 1–8.

[12] S. E. Olive and P. L. Schuck, “The Variability of Loudspeaker Sound Quality Amoung Four Domestic-Sized Rooms,” in *99th Audio Engineering Society Convention*, 1995.

[13] T. Hiekkanen, A. Mäkivirta, and M. Karjalainen, “Virtualized Listening Tests for Loudspeakers,” *J. Audio Eng. Soc.*, vol. 57, no. 4, pp. 237–252, 2009.

[14] H. Wierstorf, “Perceptual Assessment of Sound Field Synthesis,” Technische Universität Berlin, 2014.

[15] F. Melchior, D. Marston, C. Pike, D. Satongar, and Y. W. Lam, “A Library of Binaural Room Impulse Responses and Sound Scenes for Evaluation of Spatial Audio Systems,” in *40th Annual German Congress on Acoustics*, 2014.

[16] M. Geier, J. Ahrens, and S. Spors, “The SoundScape Renderer: A Unified Spatial Audio Reproduction Framework for Arbitrary Rendering Methods,” in *124th Audio Engineering Society Convention*, 2008.

[17] A. Lindau, L. Kosanke, and S. Weinzierl, “Perceptual Evaluation of Model- and Signal-Based Predictors of the Mixing Time in Binaural Room Impulse Responses *,” *J. Audio Eng. Soc.*, vol. 60, 2012.

[18] C. Pike, F. Melchior, and T. Tew, “Assessing the Plausibility of Non-Individualised Dynamic Binaural Synthesis in a Small Room,” in *55th AES International Conference*, 2014, pp. 1–8.

[19] B. Masiero and J. Fels, “Perceptually Robust Headphone Equalization for Binaural Reproduction,” in *130th Audio Engineering Society Convention*, 2011, pp. 1–7.

[20] A. M. Salomons, “Coloration and Binaural Decoloration of Sound Due to Reflections,” Technische Universiteit Delft, 1995.

[21] B. S. Atal and M. R. Schroeder, “Perception of Coloration in Filtered Gaussian Noise - Short-time Analysis by the Ear,” in *Fourth International Cogress on Acoustics*, 1962.

[22] P. M. Zurek, “Measurements of binaural echo suppression,” *J. Acoust. Soc. Am.*, vol. 66, no. 6, pp. 1750–1757, 1979.

[23] F. A. Bilsen and R. J. Ritsma, “Some Parameters Influencing the Perceptibility of Pitch,” *J. Acoust. Soc. Am.*, vol. 47, no. 2, pp. 469–475, 1970.

[24] S. Olive and F. Toole, “The Detection of Reflections in Typical Rooms,” *J. Audio Eng. Soc.*, vol. 37, no. 7/8, pp. 539–553, 1989.

[25] J. M. Buchholz, “Characterizing the monaural and binaural processes underlying reflection masking,” *Hear. Res.*, vol. 232, pp. 52–66, 2007.

[26] P. L. Søndergaard, J. F. Culling, T. Dau, N. Le Goff, M. L. Jepsen, P. Majdak, and H. Wierstorf, “Towards a binaural modelling toolbox,” *Forum Acoust.*, 2011.